First-Generation Transgenic Plants and Statistics

Nap, Jan-Peter; Keizer, Paul; Jansen, Ritsert

Published in:
Plant Molecular Biology Reporter

DOI:
10.1007/BF02670473

Publication date:
1993

Plant Molecular Biology Reporter
Volume 11(2) 1993

pages 156–164

Reviews

First-Generation Transgenic Plants and Statistics

Jan-Peter Nap, Paul Keizer, and Ritsert Jansen

(JPN) Department of Molecular Biology and (PK, RJ) Population Biology
Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, The Netherlands

Key Words: transgenic plants, normal distribution, distribution-free inference, gene expression

Abstract: The statistical analyses of populations of first-generation transgenic plants are commonly based on mean and variance and generally require a test of normality. Since in many cases the assumptions of normality are not met, analyses can result in erroneous conclusions. Transformation of data to normality, the use of other distributions, or distribution-free statistical tests should then be used to obtain valid conclusions from these populations.

From the point of view of statistics, recombinant DNA research appears fairly simple. A particular cloning is either OK or not and for the latter you have ample checks available. The moment you transfer a construct to plants, trouble begins. You will find an immense variation in what your construct is doing in these plants. The variation as result of the so-called *position effect* is severely hampering the analyses of constructs in transgenic plants. A question commonly addressed in plant molecular biology is whether and to what extent constructs differ in activity. This way enhancers, silencers and/or sequences involved in tissue-specificity of gene expression are identified. In these analyses, regularly the very basics of statistics is violated. The result is inappropriate data presentation, doubtful data interpretation and sometimes completely erroneous conclusions. This is unnecessary when plant molecular biologists become more familiar with what statistics has to offer.

Abreviation: GUS, β-glucuronidase

Stephen Hawking once claimed that every formula halves the number of readers; we will refrain, therefore, from presenting formulae. We apologize to the mathematicians are be offended by the lack of any mathematical rigor, but please realize that the target audience is not those who can follow Bain and Engelhardt (1989), McCullagh and Nelder (1989), or Hall (1992).

In Search of Normality

Most pocket calculators and all statistical software packages allow easy calculation of mean, variance, (adjusted) standard deviation and standard error. This is probably the reason that these characteristics of the data are generally presented and used for the analysis of constructs in first-generation transgenic plants. Although mean and variance can be calculated for any data, you should note that their interpretation depends on the particular distribution of the data.

The best-known and most used distribution function is the bell-shaped 'normal' distribution. The implication of *normal* is quite unfortunate because the distribution may not be as common as one would hope or like. A normal distribution is fully specified by the two parameters mean and variance (or standard deviation). These two parameters are thus a sufficient and appropriate description of a normal population. Conversely, *only* in the case of normality does it makes sense to describe data simply by mean and variance *only* in the case of normality. Implicit in most analyses of gene expression in first-generation transgenic plants, therefore, is the assumption of normality. Fig. 1 shows the consequences. The histogram gives the number of plants with a certain level of gene expression (in this case, GUS activity), the solid line is the normal distribution assumed in this population by calculating the mean and the variance of the GUS activity data. In other words, by reporting your data set as a mean and variance or standard deviation, you *suggest* that your sample is represented by the solid line, while *in reality* it is the histogram. The high probability of obtaining a negative value demonstrates how obviously wrong this is. The statistics of the normal distribution is by far the most advanced, powerful and efficient available (see Steel and Torrie, 1980; Sokal and Rohlf, 1981). Therefore, it is worthwhile to test whether your data set may reasonably be supposed to come from a normal distribution. A quick (but dirty) indication is whether the mean is equal to the median, the median being the midmost value of the data. In case of an even number of data, the median is the midpoint between the midmost two.

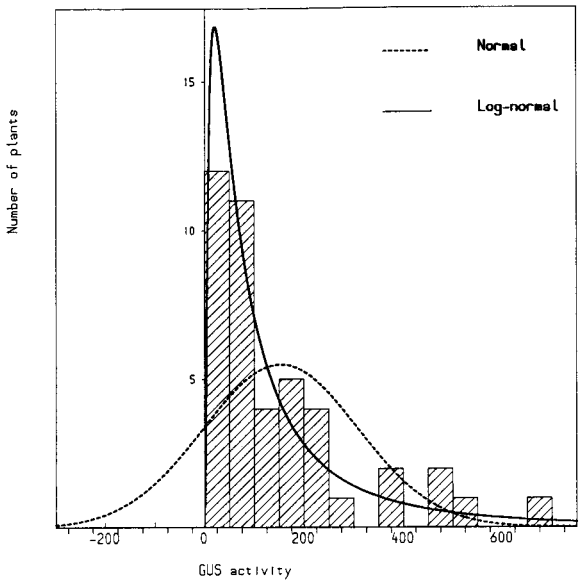

Fig. 1. Data transformation may result in a normal distribution of gene expression in populations of first-generation transgenic plants. The striped bars show the frequency distribution of GUS gene activity (expressed in pmol·min^{-1}·µg protein^{-1}) in a population of 43 first-generation transgenic tobacco plants all carrying the same promoter-GUS fusion. The distribution fitted to these data is given as solid line. Activities are: 5.60 6.40 7.71 16.05 17.17 17.93 19.31 29.53 31.67 34.57 35.79 49.68 52.19 53.31 69.78 73.03 80.37 82.83 87.97 88.36 91.58 95.97 99.06 110.66 111.99 136.81 144.90 152.34 166.89 173.89 175.48 183.89 203.76 213.85 221.74 245.60 269.76 350.37 383.28 480.06 481.88 524.61 679.25 pmol·min^{-1}·µg protein^{-1}. When GUS activity is used directly for fitting, the solid line (—) is obtained. Statistical characteristics of this population are: mean = 152; variance = 23936; SD(n-1) = 157; standard error of mean = 24; median = 96; MAD = 71; Lilliefors' statistic = 0.183 (P for normality < 0.01%); Shapiro-Wilk statistic = 0.81 (P for normality < 0.01%); Lagrange multiplier statistic = 27.8 (P for normality < 0.01%). When the natural logarithm of the GUS activity is used for fitting, and the resulting fit is re-transformed to the linear scale of measurement, the dotted line (···) is obtained. Statistical characteristics of the transformed values of this population: mean = 4.5; variance = 1.40; SD(n-1) = 1.20; standard error of mean = 0.18; median = 4.56; MAD = 0.75; Lilliefors' statistic = 0.102 (P for normality > 99%); Shapiro-Wilk statistic = 0.958 (P for normality > 80%); Lagrange multiplier statistic = 2.01 (P for normality > 95%).

Various tests are available to test the assumption of normality, *e.g.*, the Lilliefors' test (Conover, 1980; Sprent, 1989), the Shapiro-Wilk test (Conover, 1980), and the Lagrange multiplier test (Jarque and Bera, 1987). The latter test is least known, but claimed to be the most powerful in a variety of cases (Jarque and Bera, 1987). As these tests are based upon different principles, you could use all three of them. All three tests confirm that the data set shown in Fig. 1 has *not* a normal distribution. It is important to realize that these tests do *not* allow you to decide whether your data set *is* normal. In fact, no finite data set will ever be normal. These tests indicate, however, whether normality is a reasonable approximation of the true but unknown distribution of your data, so that you are allowed to use the powerful statistics of the normal distribution. To facilitate the interpretation of your data by others, you should cite the tests performed and the test statistics obtained.

Outliers, which are the extreme data points, are devastating for any test of normality, notably so in relatively small data sets. Because the Lagrange multiplier test uses the actual data and not some representation derived from them, it is more sensitive to outliers than the other two. To see whether the assumption of normality is violated by one or a limited number of outliers, you can 'trim' your data set, that is discard extremes on either side, and analyze again. Whether or not you should discard data in your final analyses depends on the grounds you have for discarding, *e.g.* Southern blotting shows the gene to be absent, and on your integrity.

In Case of Non-Normality

We are quite confident that you will find that the gene expression in your population of first-generation transgenic plants is *not* following a normal distribution (see Fig. 1). You now have four possibilities for analysis. You can either ignore it, try to transform your data to a normal distribution, use distribution-free tests, or try to find a model to describe your data in terms of other types of distributions.

Ignoring. This means essentially that you don't intend to perform any normality test. Editors or reviewers do not seem to bother very much, so why should you? After all, ignoring out of ignorance cannot be blamed, can it? But after reading this far, you are no longer innocent! Say you report construct I: mean = 3322, SD = 3348, n = 17; and construct II: mean = 7127, SD = 5693, n = 18. (No reference out of courtesy). What does it

mean? It depends, unfortunately. The standard two-sample t-test (Parker, 1979) shows the difference to be significant ($0.02 < P < 0.05$), but assumes that both populations have a normal distribution *and* that both populations have the same variance. The latter is clearly untrue, the first should be tested, and, in our experience, is also likely to be untrue. Without more information, the conclusion that construct II is approximately two times more active than construct I is, to say the least, doubtful. Unfortunately, this kind of conclusion is fairly common in plant molecular biology. Depending on the distribution of your actual data, notably the presence of putative outliers, it is even possible that construct I should be considered the 'better' one.

Data transformation. Believe it or not, but after you have transformed *Escherichia coli, Agrobacterium tumefaciens* and your plant of interest, you can now try to transform your data. As the statistics of the normal distribution is the most powerful, you can try to put your data in a form that *is* normally distributed. Common transformations are the inversion, square root, logarithmic and (for proportions and percentages) arcsine and logit transformation. Various combinations and improvements of these transformations have been found successful. An attractive approach for finding the optimal transformation is the Box-Cox transformation. This is a family of power transformations including some of the standard transformations mentioned above (Haaland, 1989). Unfortunately the Box-Cox transformation method is missing from several existing statistical software packages.

The transformed data can be evaluated for normality by the same tests mentioned above. Fig. 1 shows the approximation of the data set shown before (Fig. 1) after a logarithmic transformation. The fit after logarithmic transformation is improved considerably compared to the fit of the data on the scale of measurement. All tests mentioned above confirm that the logarithms of the data represent an acceptable approximation of a normal distribution. Generally, when the number of 'zeros' in your data set is not too large, a logarithmic transformation in which a small number is added to the zeros may bring normality in sight.

Once you have decided that a particular data transformation gives a reasonable approximation of normality, you should perform all analyses in that scale of measurement. For example, if you are using the logarithmic transformation and you perform three replicates on a particular measurement (*e.g.*, GUS activity), you should take the mean of the logarithms of the three measurements instead of the logarithm of the

mean of the three. You will find then that the results are (slightly) more consistent. When you report, you can back transform the results to the measurement scale as was done in Fig. 1.

Data transformation may, at first, look like fudging. It is not. Some variables just happen to have a non-linear scale. Every time you measure the pH of a solution, you use a linear scale as result of the transformation of a logarithmic variable. Gene expression may be such a variable. Be aware, however, that , for example, in comparing two populations of transgenic plants the *same* transformation should be appropriate and used for both populations.

Distribution-free tests. If you have bad luck, all attempts to transform your data to a normal distribution may fail. Generally, but especially in this case, distribution-free tests are a valuable extension of the statistical tools available. As the name already suggests, distribution-free tests assume no underlying (normal) distribution of your data. In distribution-free statistics, most often the rank is used, that is the relative position of an observation in the total population of observations. Outliers, therefore, have little, if any, effect. In addition, these tests are valid for observations in counts, or categories. In general, the tests are relatively simple, and can have a remarkable power. This is especially the case when the assumption of normality is not met or when the sample size is fairly small, as is the case for populations of first-generation transgenic plants. Distribution-free tests, however, are *not* assumption-free (Conover, 1980). The various distribution-free tests have much to offer to the plant molecular biologist. In fact, the Lilliefors' and Shapiro-Wilk tests mentioned above to check normality belong to the distribution-free tests.

We argued above that mean and variance need not be very appropriate, hence informative, descriptions of your population of transgenic plants. The measure most suitable to describe the location of an unknown distribution probably is the median. A description not very commonly used, but certainly worth while, to describe the shape of such an distribution is the Median Absolute Deviation or MAD (Sprent, 1989). Note how terminology is discriminating this approach: MAD versus normal! MAD is the median of the set of absolute differences between each observation *and* the median of all observations. Generally, reporting the median and MAD of a non-normal population of transgenic plants is likely to be more informative than reporting the mean and variance.

For most statistical data analyses, distribution-free methods are available. Unfortunately, the tests can be performed in slightly different ways,

so the plant molecular biology community should agree upon the way these tests are applied and reported. We will not go into details of the various tests, the names of all statisticians involved probably being as deterrent as formulae would be. The one test worth mentioning has even three names attached to it. The Wilcoxon-Mann-Whitney test is extremely useful for determining whether one construct is performing better than another (Sprent, 1989). Very readable and useful guides to the practical applications of distribution-free methods are the books by Neave and Worthington (1988) and Sprent (1989), while some more mathematics is used in Conover (1980). More advanced, but quite digestible is the book by Hollander and Wolfe (1973).

Distribution-free methods are traditionally called 'non-parametric.' This name immediately indicates a major drawback, namely that you are able to analyze statistically, but you are not able to quantify any effect. The Wilcoxon-Mann-Whitney test can indicate that a particular construct is significantly better than another, but you are not able to say *how much* better. Also the discriminative power of distribution-free tests may be somewhat less than the normal distribution statistics, but only when the latter is valid. However, the advent of computers holds great promises for both parameter and confidence estimation in combination with these methods. Recurrent re-samplings using either jackknife or bootstrap approaches (Efron, 1982; Hall, 1992) are likely to contribute to a more accurate description of populations, but more research is required to establish in what situations these methods are accurate. Ask your local statistician. The research questions asked today in the analyses of gene expression in first-generation transgenic plants do not seem to have reached the level of sophistication where these advanced distribution-free techniques may be of use.

Modeling. Depending on your particular research question, quantification may be required also when you are not able to transform your data to the normal distribution. In addition to this normal distribution, statisticians have described and studied numerous other distributions (see Bain and Engelhardt, 1989), such as the Gamma or Weibull distribution. Some of them have a clear physical or biological interpretation, others are merely mathematical descriptions of observed data. You can try to fit your data to one of these known distributions, or assume mixtures of them. Also the Generalized Linear Model approach may be useful to analyze your data (see McCullagh and Nelder, 1989). Once you have found a proper model to describe your data set, you can use this model for analysis and quantification. Modeling probably gets the most

information out of your data and allows meaningful quantification of effects. In general, however, this requires advanced mathematics and computing that is far beyond most plant molecular biologists. While all statistical tools give better results in the hands of an experienced user, modeling tends to require more experience and subtlety than other techniques. Again, ask your local statistician. He or she will probably tell you that you do have not enough data to make modeling meaningful. The size of populations of first-generation transgenic plants is likely to remain small because of practical, economical and possibly regulatory reasons. Therefore, we tend to think that this advanced statistics is not yet appropriate for the comparative analyses of constructs in first-generation transgenic plants. But this opinion may well be caused by the fact that until now no fanatic modeling statistician has had a close look at the distribution of gene expression in first-generation transgenic plants.

Plant Molecular Statistics?

Plant molecular biologists need not become statisticians and this is by far a plea for establishing the field of plant molecular statistics. However, some rudiments of statistics may be helpful in reporting on first-generation transgenic plants. When the statistics of the normal distribution, mean and variance or others, are used, appropriate tests for normality, either before or after data transformation, should be performed. A logarithmic transformation may be quite helpful. In case of non-normality, we suggest that instead of, or in addition to, mean and variance, median and MAD are reported. Non-parametric tests for various valid research questions are available and attractive. These are worth using more often. As the numbers of first-generation transgenic plants will remain relatively small, the data themselves should preferably be reported to allow reevaluation by others. Greater care in reporting data on transgenic plants is likely to improve the validity of the conclusions drawn considerably.

Acknowledgments. We would like to thank Nelleke Kreike, Peter Metz, Ludmila Mlynarova, Willem Stiekema and Fred van Eeuwijk for sharpening our thoughts on the subject. This work was motivated by EEC-BRIDGE grant CT-910298 on the biosafety of transgenic crops.

References

Bain, L.J., and M. Engelhardt. 1989. *Introduction to Probability and Mathematical Statistics.* PWS-Kent Publishing Co, Boston.

Conover, W.J. 1980. *Practical Nonparametric Statistics,* 2nd Ed. John Wiley & Sons, New York.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans.* Soc. Indust. Appl. Math., Philadelphia, Pennsylvania.

Haaland, P.D. 1989. *Experimental Design in Biotechnology*, Marcel Dekker, Inc., New York.
Hall, P. 1992. *The Bootstrap and Edgeworth Expansion*. Springer Verlag, New York.
Hollander, M., and D.A. Wolfe. 1973. *Nonparametric Statistical Methods*. John Wiley & Sons, New York.
Jarque, C.M., and A.K. Bera. 1987. A test for normality of observations and regression residuals. Int. Stat. Rev. 55:163-172.
McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models*, 2nd Ed. Chapman and Hall, London.
Neave, H.R., and P.L. Worthington. 1988. *Distribution-free Tests*. Unwin Hyman, London.
Parker, R.E. 1979. *Introductory Statistics for Biology*. Edward Arnold, London.
Sokal, R.R. and F.J. Rohlf. 1981. *Biometry*, 2nd Ed. W.H. Freeman and Company, New York.
Sprent, P. 1989. *Applied Nonparametric Statistical Methods*. Chapman and Hall, London.
Steel, R.G.D., and J.H. Torrie. 1980. *Principles and Procedures of Statistics. A Biometrical Approach*. McGraw-Hill Kogakusha, Tokyo.